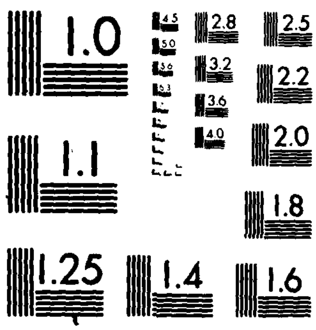
AD-A097 537 CLEMSON UNIV SC DEPT OF MATHEMATICAL SCIENCES F/G 12/1
PARAMETRIC COST METHODOLOGY.(U)
SEP 80 A S KORKOTSIDES, K T WALLENIUS N00014-75-C-0451
UNCLASSIFIED N121 NL

END
DATE
FILMED
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

DISTRIBUTION STATEMENT A

Approved for public release;
Distribution Unlimited

LEVEL II 6

6 PARAMETRIC COST METHODOLOGY.

by

10 A. S. Korkotsides
[redacted]
K. T. Wallenius
Clemson University

DEPARTMENT OF MATHEMATICAL SCIENCES

9 TECHNICAL REP[redacted]RT #353

12 25

11 SEP[redacted]80

REPORT N121, TR-353
14

This report forms the basis of an invited paper to be presented at the I.E.E.E. 1981 Annual Reliability and Maintainability Symposium, January 27-29, 1981.

15

This work was supported in part by the Office of Naval Research under Contract N00014-75-C-0451

I. BACKGROUND

Accurate cost estimation of alternative weapons systems is essential for long range military planning and budgeting decisions. The choice among competing systems is based on trade-offs between performance parameters and mission requirements with an ever increasing importance being placed on cost. In fact, with the issuing of Department of Defense (DoD) Instruction 5000.1, cost has been upgraded to a principle design parameter. That directive defines specific "Design to Cost" policies during requirements formulation in determining which system is most cost effective.

Traditionally, weapon system cost estimates have been prepared using Industrial Engineering (I.E.) techniques. These techniques involved detailed studies of the operations and materials required to produce the new system. The cost estimate frequently required several thousand hours to produce with volumes of supporting documentation. Small changes in design often necessitate extensive revision of these estimates. In spite of all the time and effort involved in preparing these estimates, their accuracy leaves much to be desired. This is evidenced by the large cost overuns cited by the annual General Accounting Office (GAO) reports to Congress. In 1972, for example, the GAO reported that the Navy had experienced a cost growth of $19 billion on 24 weapon systems in FY 1971. Approximately 15% of this cost growth was attributed to poor initial cost estimates for the weapon systems. The report went on to make the following recommendation:

> "Develop and implement DoD wide guidance for consistent and effective cost estimating procedures and practices particularly with regard to, ... an effective independent review of cost estimates."

This report and other high level directives implementing its recommendations have resulted in increased interest and reliance on a statistical approach called <u>parametric cost estimation</u> defined by Baker (2) as

> "An estimate which predicts costs by means of explanatory variables such as performance characteristics, physical characteristics, and characteristics relevant to the development process, as derived from experience on logically related systems."

The construction and use of "cost estimating relationships" (CER's) is the basis for making independent parametric cost estimates. CER's are mathematical equations which relate system costs to various explanatory variables. They are based on the premise that the cost of a weapons sytem is logically related in a quantifiable way to the system's physical and performance characteristics. They are most generally derived through statistical regression analysis of historical cost data. These techniques are described in (14). Some examples of their use appear in (9), (10), (15), and (17).

The parametric approach has some distinct advantages and disadvantages compared to I.E. methodology. On the plus side are:

1. Parametric cost estimates can be developed during the concept formulation stage of the acquisition process before detailed engineering plans are available. These early cost estimates can be used to:

   (a) Identify possible cost/performance trade-offs in the design effort.

   (b) Provide a basis for cost/effectiveness review of performance specifications.

   (c) Provide information useful in the ranking of competing alternatives.

   (d) Suggest a need for identifying and considering new alternatives.

2. Historical cost data incorporates system development setbacks such as engineering and design specification changes and other items that are not identifiable at the time of design. Industrial engineering estimates tend to be optimistic in that they don't allow for unforeseen problems. Unexpected engineering or design changes usually bring about unexpected increases in system cost. Cost estimating relationships based on historical data will incorporate some of these unknowns into the cost estimate.

Along with the cited advantages of parametric cost estimation, there are some troublesome aspects. Assumptions, subjective assessments and choice of methodolgy can lead different analysts to different estimates. This kind of ambiguity seems unavoidable in empirical model building studies and is the basis for some critcism. While there are general guidelines as to what constitutes sound statistical practice, there is rarely a universally accepted "best" approach to modeling and prediction.

This paper focuses on the role of analogy in empirical model building. A measure of analogy between systems of the same general type is proposed and its properties studied. A new criteria for variable selection based on this measure is explored and a parametric cost estimation example is given.

II. THE ROLE OF ANALOGY IN PREDICTION

The phrase "logically related system" in the cited definition of parametric cost estimation is subject to all kinds of interpretation and degrees of relation. Certainly there is no historical system identical in all respects to the object system (the system whose cost we wish to predict) else the problem would not exist. At the other extreme, all military systems are "logically related" in (at least) the sense that they are military systems. Message carrying pidgeons, air-to-air missiles, jet aircraft and frizbees are "logically related" in that they all fly. Obviously, the analyst must take into account the degree of analogy between each system (which is a candidate for the historical data base) and the object system. Analogy, according to Webster, is "a partial similarity between like features of two things on which a comparison may be based." How does one measure the degree of analogy between "logically related" systems and how can one exploit these partial similarities in predicting the cost of an object system?

Having gathered and adjusted historical data on systems judged more-or-less analogous to a proposed system whose cost is to be estimated, the analyst proceeds with the task of developing a "best" CER. This involves selecting the form of the CER, deciding which of the system variables (performance characteristics, design specifications, etc.) to include as predictor variables, and assessing the precision of the estimate. In parametric cost estimation, this is usually done through the use of multiple regression and some standard variable selection criterion such as maximizing adjusted R^2 (which is equivalent to minimizing mean square error [MSE]), maximizing F, using Mallow's C_p, etc.

All of these techniques share two properties: (1) For any fixed number of variables in the prediction equation, the optimal set of variables is that set which minimizes the MSE; (2) They all ignore the values of the variables of the system whose cost is being estimated. The first of these properties is reasonable but myopic when the object is prediction. The second property seems contrary to common sense.

Suppose there are n systems in the historical data base. Associated with the i^{th} such system is a cost Y_i and values of p (candidate) predictor variables $X_{i1},\ldots,X_{ij},\ldots,X_{ip}$. Let $\underset{\sim}{Y}' = (Y_1\ldots Y_i\ldots Y_n)$ and $\underset{\sim}{X}_j' = (X_{1j}\ldots X_{ij}\ldots X_{nj})$, $j = 1,2,\ldots,p$ denote these historical costs and system characteristics. Furthermore, let

$$\overline{X}_j = \frac{1}{n}\sum_{k=1}^{n} X_{kj} \qquad \text{and}$$

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n} (X_{ki}-\overline{X}_i)(X_{kj}-\overline{X}_j)$$

denote the sample means and covariances. Denoting the values of the proposed system by lower case letters, we wish to predict its cost y by exploiting the predictive ability its characteristics $X' = (x_1,\ldots,x_j,\ldots,x_p)$. This predictive ability is inferred from the apparent relation between historical costs and characteristics <u>and</u> the degree of analogy between the proposed system and these historical data. How analogous is this proposed system to the historical data?

(a) Marginal comparisons: Analogy on a single dimension is straightforward. A statistic commonly used as a nonnegative distance index is simply the square of the standardized distance between x_j and the mean of the X_{ij}'s, namely

$$M_j = \left(\frac{x_j-\overline{X}_j}{s_j}\right)^2$$

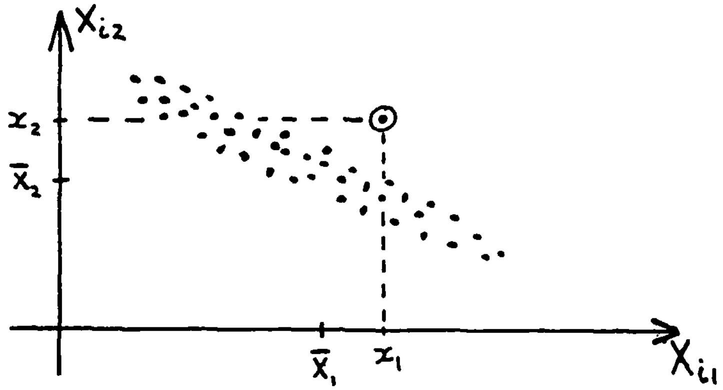
where s_j is defined as $(s_{jj})^{\frac{1}{2}}$. Large values of this statistic indicate a low degree of analogy.

(b) High dimensional comparisons: The collection of marginal indices $\{M_1,M_2,\ldots,M_p\}$ can give a very misleading impression of the overall degree of analogy. Even when M_j is small for <u>every</u> j, the proposed system can be terribly non-analogous to the historical data. A simple bivariate example will illustrate this assertion. Suppose X_1 and X_2 denote weight and maneuverability, respectively, and that x_1 and x_2 are each within one standard deviation of their respective means, i.e.,

$$M_1 = M_2 \leq 1.$$

Suppose further that, historically, heavy systems tended to be less maneuverable, i.e., $\rho_{x_1x_2} < 0$, but the proposed system is a little heavier and slightly <u>more</u> maneuverable than the average. The situation is depicted in

Figure 1



We see that (x_1, x_2) is marginally analogous to the historical data on both weight and maneuverability but not at all analogous when viewed in two dimensions. Comparing (x_1, x_2) to $(\bar{X}_1, \bar{X}_2)$ marginally ignores important relational information. The geometry in higher dimensional spaces is more difficult to grasp.

A measure of analogy which incorporates relational information was suggested in 1930 by P.C. Mahalanobis (12). He proposed

$$M = (\underline{\mu}^{(1)} - \underline{\mu}^{(2)})' \textstyle\sum^{-1} (\underline{\mu}^{(1)} - \underline{\mu}^{(2)})$$

as a measure of the distance between two multivariate populations with mean vectors $\underline{\mu}^{(1)}$ and $\underline{\mu}^{(2)}$, respectively, and common covariance matrix $\sum$. Replacing the parameter values by estimates, we obtain (in our notation)

$$\hat{M} = (\underset{\sim}{x} - \underset{\sim}{\bar{X}})' S^{-1} (\underset{\sim}{x} - \underset{\sim}{\bar{X}})$$

where $S = (s_{ij})$ and $\underset{\sim}{\bar{X}}' = (\bar{X}_1, \ldots, \bar{X}_p)$. Except for a multiplicative constant, this is Hotelling's T^2 statistic used to test that $\underset{\sim}{x}$ and the historical data came from the same population. In the previous bivariate example, it is easy to show that

$$\hat{M} = \frac{1}{1-\hat{\rho}^2}\,[M_1 - 2\hat{\rho}\,(M_1M_2)^{\frac{1}{2}} + M_2],$$

which can be arbitrarily large even when M_1 and M_2 are small. For example, with $M_1 = M_2 = \varepsilon$,

$$M = \frac{2\varepsilon}{1+\hat{\rho}} \quad \text{and} \quad \lim_{\hat{\rho}\to -1} \hat{M} = \infty.$$

In what follows, we propose Mahalanobis distance $\hat{M}$ as a measure of analogy and discuss its implications in the process of tailoring a CER to a specific object system for the purpose of predicting that system's cost. This is a distinct departure from standard procedures recommended (14), (15) and used in developing every CER with which the authors are familiar. The distinction is fundamental and goes beyond measure of analogy. The standard approach appears more oriented toward developing a cost explaining equation relating costs of a class (e.g., sonars, airframes, tanks, etc.) of historical systems to the characteristics of those systems. One need not have any specific object system in mind while developing such a general purpose descriptive equation. In fact, armed with an airframe CER based on the explanatory variable "weight", two radically different airframes of the same weight would be estimated to cost the same amount. Mallows (13) defined six potential uses of a regression equation which include (a) pure description and (b) prediction. Lindley (11) emphasizes that the technique used to develop a regression equation ought to be related to the intended use. In the present context, the intended use is prediction of the cost of a specific system so that using a CER (which was developed to describe historical relations without reference to any specific object system) to predict cost of a specific system is contrary to Lindley's recommendation and common sense.

III. EXPLOITING ANALOGY IN DEVELOPING MODELS FOR PREDICTION

As mentioned in the previous section, most standard variable selection techniques share the property that, for any given number of variables in the regression function, the optimal set is that set which minimizes residual mean square error (or, equivalently, maximizes r^2). The objective system may be rather nonanalogous to the historical data (large $\hat{M}$) when we consider the subset of variables identified as "optimal" by the criteria used to develope the CER. Often, there are several k-variable models which come close to the "optimum" in terms of r^2 and other measures of model aptness based on residual analysis. In these cases, by using a slightly suboptimal set of prediction variables (slight decrease in r^2) it may be possible to substantially improve the degree of analogy (decrease in $\hat{M}$). What is the role of analogy in prediction and how can one evaluate the trade-off of fit for analogy?

The width of the prediction interval at the point corresponding to the objective system is a numeraire which seems like a reasonable basis for choosing between alternative models. We shall consider a monotone function of the width for simplicity, namely, the square of the half-width, viz.

$$W = F_{1-\frac{\alpha}{2};1,n-k-1} * MSE * \left(\frac{\hat{M}}{n-1} + \frac{n+1}{n}\right),$$

where $F_{1-\frac{\alpha}{2};1,n-k-1}$ is the $(1-\frac{\alpha}{2})^{th}$ fractile of an F distribution with 1 and n-k-1 degrees of freedom. This measure W combines "fit" (MSE) and "degree of analogy" (M) with a factor F which penalizes for using too many variables (increasing k) or excluding points from the data base (decreasing n). In this form, the role of analogy, as measured by Mahalanobis distance, is evident. It enters as a term in the multiplier $\left(\frac{\hat{M}}{n-1} + \frac{n+1}{n}\right)$ of MSE. Failure to consider this factor in selecting a CER could have a marked effect on predictor precision as measured by prediction interval width. We hasten to point out that the W-criterion is not being suggested as a universally "best" variable selection procedure. Rather, it should be viewed as a device to reduce the 2^p candidate models to a more manageable number so that detailed residual analysis is economically feasible. Models which appear attractive under criteria based on maximizing R^2 but are far from W-optimal obviously involve relatively large extrapolation at the point corresponding to the object system. With extrapolation comes concern that the model which seems to fit the historical data best may not be valid at the point under prediction.

IV. PROPERTIES OF W

While the quantity W is related to a $100(1-\alpha)\%$ prediction interval, it is not true that $\hat{y}(x) \pm W^{\frac{1}{2}}$ covers $y(x)$ with probability $1-\alpha$ when the same data is used to select the model and compute the prediction interval. The confidence coefficient resulting from using the model with the smallest W will, in general, be less than $1-\alpha$. This drawback is not limited to the W-criterion but is a general problem associated with empirical model building and subsequent inference using the same data.

It is well known (for example, see (5) and (16)) that choosing the k-variable model which maximizes R^2 has an inflationary effect on the distribution of R^2, invalidating the usual significance tests. Distribution-free exact tests have recently been developed by Edwards and Wallenius (5) to deal with this problem. Distributional properties of W are not well understood unless model building and inference are made on independent data sets, a luxury of sample size not commonly encountered in parametric cost estimation. A monotonicity result which is useful in variable selection algorithms and decisions concerning inclusion or exclusion of points in the historical data base is given in the following

<u>THEOREM</u>: Let $\hat{M}_k$ denote the Mahalanobis distance between $\underset{\sim}{x}_k' = (x_1,\ldots,x_k)$ and $\underset{\sim}{\overline{X}}_k' = (\overline{X}_1,\ldots,\overline{X}_k)$ for $k = 1,\ldots,p$ and let $\Delta_j = \hat{M}_{j+1}-\hat{M}_j$.

Then $\Delta_j = \dfrac{(x_{j+1}-\hat{x}_{j+1})^2}{S_{22}(1-r_j^2)}$ where

$$r_j^2 = S_{21}S_{11}^{-1}S_{12}/S_{22}, \quad \hat{x}_{j+1} = \overline{X}_{j+1}+S_{21}S_{11}^{-1}(\underset{\sim}{x}_j-\underset{\sim}{\overline{X}}_j)$$

and $S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$ is the partitioned covariance matrix of

$X_1,X_2,\ldots,X_{j+1}$ with S_{22} being a scalar.

A proof of the theorem is given in (8). Since $0 \leq r^2 < 1$ and S_{22} is the sample variance of X_{j+1}, it follows that $\Delta_j \geq 0$ with equality if, and only if, $x_{j+1} = \hat{x}_{j+1}$, that is, if x_{j+1} lies on the hyperplane defined by the regression of X_{j+1} on the variables $(X_1,\ldots,X_j)$. Thus, augmenting a specific j-variable cannot decrease Mahalanobis distance (i.e., increase analogy) so that, using the W-criterion, a variable will be added to a given model only if, by so doing, the decrease in MSE is large enough to offset increases in $\hat{M}$ and F. This observation explains why W-optimal models tend to be more parsimonious than MSE optimal models, a desireable property in view of the commonly held opinion among satisticians that the minimum MSE criterion frequently results in overfitting.

Besides providing this kind of insight into the effect on W of adding an explanatory variable to a given model, the monotonicity theorem is useful in developing a branch and bound algorithm similar to that of Furnival (6) which makes it possible to efficiently identify models with small W's without computing all 2^p-1 possible regressions. These computational aspects are discussed in (8) along with a treatment of various suboptimal search algorithms (e.g., forward selection, backward selection, etc.)

The monotonicity theorem is also useful in dealing with questions about setting aside points in the historical data base. It is quite plausible that inclusion of one or more systems in the data base which are rather nonanalogous to the object system and/or the other reference systems could have a detrimental effect on prediction. The influence of individual observations or groups of observations on various quantities of interest in a statistical analysis has received considerable attention in recent literature (3), (4), (7), etc. Most of this work is concerned with identifying outliers and assessing the effect of setting them aside on estimates of model parameters. Since the focus of our work is on prediction, these earlier studies are not directly applicable. Space limitations here do not allow for treatment of questions on the role of analogy in choosing the data base. The interested reader is referred to (8) for further information.

V. APPLICATION

Data that formed the basis for an actual parametric cost estimation study (18) was used to compare the performance of the W-criterion with other standard and not-so-standard variable selection techniques. The data consisted of costs and 12 design/performance characteristics of 23 single engine jet interceptors built between 1947 and 1969. The object of the study was to predict costs of the F-14 and F-15 aircraft. As is common practice, see (10), (14), and (15) for example, the data were log-transformed prior to analysis. (Our analysis confirmed that regressions were more nearly linear and residuals better behaved using the transformed data.) Cost was defined as "total flyaway cost (adjusted to 1972 constant dollars x 10^6) of the 100^{th} production aircraft." The candidate predictor variables were

X_1 = Wing Loading Ratio

X_2 = Aspect Ratio

X_3 = Full to Empty Weight Ratio

X_4 = Thickness-to-Cord Ratio

X_5 = Lift to Drag Ratio

X_6 = Total Avionics Input Power in kva

X_7 = Maximum Speed in knots (Clean, Combat Weight)

X_8 = Weight Empty in lbs

X_9 = Rate of Climb in ft/min, (sea level, combat weight and power)

X_{10} = Combat Ceiling in feet

X_{11} = Ferry Range in nautical miles

x_{12} = Sea Level Static Thrust (max) in lbs.

The criteria compared were Minimum MSE, Minimum C_K, Maximum F, Maximum R^2 (subject to a 1% "elbow rule"[1]), Minimum Mean Square Error of Prediction (MSEP[2]), and Minimum W (nominal 95% prediction interval). The procedure followed consisted of setting aside each aircraft, finding the best model under each of the 6 criteria based on the remaining 22 data points and predicting the cost of the deleted aircraft using these "optimal" models. While selecting a "best" CER based on any single variable selection criterion is a myopic data analytic practice, it is a necessary evil if the relative performance of different criteria are to be compared over a given data set. Absolute errors (differences between actual cost and predicted cost) and Minimum W gain (difference in Minimum W absolute error and the absolute errors of other optimal models) were computed and are displayed in Table 1. A gain denoted by "0" indicates the optimal model selected by a given criterion coincided with the Minimum W model while "00" indicates different models but a gain of less than $.5 \times 10^{-2}$.

---

[1] Since R^2 is nondecreasing as variables are added to the prediction equation, a stopping rule is needed. It is rather common to stop adding variables when the marginal increase in R^2 is small. We defined "small" to mean "no more than 1% increase in R^2 could be achieved by adding more variables."

[2] Minimum MSEP, proposed in reference (1), is the only other selection criterion besides minimum W which takes into account the point at which prediction is to be made. While the concept seems reasonable at first glance, its successful application requires a good estimate of the bias in each submodel. This logically requires knowledge (or a good estimate) of the true $E(y|x)$. But if this quantity were known, the prediction problem would be solved. The results of numerical work confirms our theoretical skepticism of Allen's approach. Detailed criticism can be found in (8).

Table 1: Absolute Errors and Minimum W Gain (x 10^2)

	Min MSE		Min C_k		Max F		R^2		MSEP		Min W
A/C	|Err|	Gain	|Err|	Gain	|Err|	Gain	|Err|	Gain	|Err|	Gain	|Err|
F-80	106	54	149	97	149	97	149	97	104	52	52
FH-1	5	0	6	-1	32	27	6	1	12	7	5
F2H-1	99	10	89	0	85	-4	83	-6	97	8	89
F7U-1	9	-15	24	0	26	2	24	0	17	-7	24
F-84E	82	55	27	0	13	-14	27	0	64	37	27
F3D-1	9	0	9	00	28	19	7	-2	19	10	9
F-86H	32	6	25	-1	81	55	25	-1	52	26	26
F9F-8	5	-54	4	-55	62	3	4	-55	4	-55	59
F4D-1	11	11	00	0	28	28	0	00	0	0	0
F3H-1N	55	43	12	0	48	36	12	0	37	25	12
F-102A	74	64	2	-8	38	28	10	0	46	36	10
F-100D	17	0	19	2	42	25	19	2	14	-3	17
FJ-4	18	-8	26	0	23	-3	26	0	1	-25	26
F-104A	71	24	38	-9	42	-5	38	-9	84	37	47
F11F-1	2	-8	10	0	28	19	10	0	9	-1	10
F-105B	45	6	39	0	42	3	39	0	45	6	39
F-101C	4	-1	5	0	58	53	5	0	6	-1	5
F-106B	17	12	5	0	12	7	5	0	18	13	5
F-4B	12	0	13	1	2	-10	13	1	9	-3	12
F-5A	61	16	45	0	84	39	45	0	66	21	45
F-4J	19	0	17	-2	31	12	17	-2	10	-9	19
F-111A	109	87	22	0	58	36	30	8	80	58	22
F-8E	24	1	18	-5	40	17	18	-5	25	2	23
MEAN	38.5	13.2	26.2	0.9	45.7	20.4	26.6	1.3	35.6	10.2	25.3
ST. D.	35.8	30.2	33.0	23.9	31.6	25.5	32.3	23.9	32.6	24.9	21.4

In the course of the analysis, 95% prediction intervals were calculated for each of the 23 costs based on the models selected by the various criteria. Their average widths and actual coverage frequencies are given in

Table 2. Average Nominal 95% Prediction Interval Widths and Frequency of Coverage in Logarithmic Units

	Min MSE	Min C_k	Max F	R^2	MSEP	Min W
Avrg. Width	1.385	1.250	1.778	1.287	2.288	1.223
Coverage	86.96	91.30	91.30	91.30	95.65	95.65

Examining tables 1 and 2, we note that the minimum W criterion had the smallest mean absolute error of all criteria tested. By design, it has the shortest prediction interval width and, quite surprisingly, was unsurpassed in the frequency with which it covered the actual cost. At least for this data set, prediction intervals seem to be well centered and tight. This experience, while certainly inconclusive, was repeated with other data sets.

Tables 3-8 are provided for comparing actual models selected by the various criteria. The selected models were ranked according to their frequency of occurrence and these ranks were used for labeling purposes. Maximum F was most parsimonious, selecting "weight empty" as the best predictor for the cost of all but two aircraft. Minimum W, minimum C_K and R^2 were often in agreement in selecting variables 2, 5, 8, and 12. As is often the case, minimum MSE seems to overfit, using up to eleven variables in eight different models. Because MSEP selected 20 different models for the 23 aircraft (none of which were selected by other criteria), these models are indicated by "×". Incidently, the model selected in the document from which the data were obtained (18), was based on variables 4, 6, and 8, a model not selected by any of our six test criteria.

TABLE 3. Minimum W Models and Aircraft Predicted

A/C	Variables X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
F-80		2	2		2			2				2
FH-1		2	2		2			2				2
F2H-1	3	3	3	3	3		3	3	3		3	
F7U-1		1			1			1				1
F-84E		1			1			1				1
F3D-1		2	2		2			2				2
F-86H		2	2		2			2				2
F9F-8		4			4			4	4			
F4D-1		1			1			1				1
F3H-1N		1			1			1				1
F-102A		1			1			1				1
F-100D		2	2		2			2				2
FJ-4		1			1			1				1
F-104A		5						5				
F11F-1		1			1			1				1
F-105B		1			1			1				1
F-101C		1			1			1				1
F-106B		1			1			1				1
F-4B		2	2		2			2				2
F-5A		2	2		2			2				2
F-4J		2	2		2			2				2
F-111A		1			1			1				1
F-8E		2	2		2			2				2

TABLE 4. Minimum MSE Models and Aircraft Predicted

A/C	Variables											
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
F-80		3		3	3			3				3
FH-1		1	1		1			1				1
F2H-1	6	6	6	6	6		6	6	6		6	
F7U-1		1	1		1			1				1
F-84E	5	5	5	5	5	5	5		5		5	
F3D-1		1	1		1			1				1
F-86H	2	2	2	2	2		2	2	2	2	2	
F9F-8		1	1		1			1				1
F4D-1		1	1		1			1				1
F3H-1N	7	7	7	7	7		7	7		7	7	
F-102A	8	8	8	8	8		8	8	8	8	8	8
F-100D		1	1		1			1				1
FJ-4		1	1		1			1				1
F-104A	2	2	2	2	2		2	2	2	2	2	
F11F-1		1	1		1			1				1
F-105B	2	2	2	2	2		2	2	2	2	2	
F-101C		1	1		1			1				1
F-106B	2	2	2	2	2		2	2	2	2	2	
F-4B		1	1		1			1				1
F-5A	2	2	2	2	2		2	2	2	2	2	
F-4J		1	1		1			1				1
F-111A	4	4	4	4	4	4	4	4			4	
F-8E	2	2	2	2	2		2	2	2	2	2	

TABLE 5. Minimum C_k Models and Aircraft Predicted

A/C	Variables											
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
F-80				3		3		3				3
FH-1		1			1			1				1
F2H-1	4	4	4	4	4		4	4	4		4	
F7U-1		1			1			1				1
F-84E		1			1			1				1
F3D-1		1			1			1				1
F-86H		1			1			1				1
F9F-8		1			1			1				1
F4D-1		1			1			1				1
F3H-1N		1			1			1				1
F-102A		2	2		2			2				2
F-100D		1			1			1				1
FJ-4		1			1			1				1
F-104A		1			1			1				1
F11F-1		1			1			1				1
F-105B		1			1			1				1
F-101C		1			1			1				1
F-106B		1			1			1				1
F-4B		1			1			1				1
F-5A		2	2		2			2				2
F-4J		1			1			1				1
F-111A		1			1			1				1
F-8E		1			1			1				1

TABLE 6. Maximum F Models and Aircraft Predicted

A/C	Variables X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
F-80				2		2		2				2
FH-1								1				
F2H-1		3	3		3			3				3
F7U-1								1				
F-84E								1				
F3D-1								1				
F-86H								1				
F9F-8								1				
F4D-1								1				
F3H-1N								1				
F-102A								1				
F-100D								1				
FJ-4								1				
F-104A								1				
F11F-1								1				
F-105B								1				
F-101C								1				
F-106B								1				
F-4B								1				
F-5A								1				
F-4J								1				
F-111A								1				
F-8E								1				

TABLE 7. R^2 Models and Aircraft Predicted

A/C	Variables											
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
F-80				3		3		3				3
FH-1		1			1			1				1
F2H-1		1			1			1				1
F7U-1		1			1			1				1
F-84E		1			1			1				1
F3D-1		1			1			1				1
F-86H		1			1			1				1
F9F-8		1			1			1				1
F4D-1		1			1			1				1
F3H-1N		1			1			1				1
F-102A		1			1			1				1
F-100D		1			1			1				1
FJ-4		1			1			1				1
F-104A		1			1			1				1
F11F-1		1			1			1				1
F-105B		1			1			1				1
F-101C		1			1			1				1
F-106B		1			1			1				1
F-4B		1			1			1				1
F-5A		2	2		2			2				2
F-4J		1			1			1				1
F-111A		2	2		2			2				2
F-8E		1			1			1				1

TABLE 8. MSEP Models and Aircraft Predicted

A/C	Variables X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
F-80	×		×		×						×	
FH-1		×			×							
F2H-1	×	×	×	×	×	×	×	×	×			
F7U-1							×	×			×	×
F-84E										×		
F3D-1					×							
F-86H			×		×				×			×
F9F-8	×		×		×							
F4D-1				×					×			
F3H-1N						×		×				
F-102A		×		×						×		
F-100D					×		×			×		
FJ-4					×					×	×	
F-104A				×			×					
F11F-1		×							×			
F-105B								×			×	
F-101C							×					
F-106B					×			×				
F-4B	×	×	×		×						×	
F-5A						×		×				
F-4J		×			×							
F-111A					×	×						×
F-8E					×							

Table 9 indicates the importance of careful data base formation with respect to the effect on prediction interval width of including an apparently nonanalogous aircraft, the F2H-1. Deleting this data point decreased the average width of W-optimal models by about 28% on the average.

In summary, the W-criterion deserves close attention as a model building device when the object of analysis is prediction at a known point in the space of predictor variables. Mahalanobis distance is a natural measure of this analogy between similar systems. Ignoring the degree of analogy between the object system and the historical data base can result in choosing a model which, in order to predict the cost of the object system, may be required to perform a large extrapolation, a dangerous practice in regression analysis.

TABLE 9. Observation Deleted and Maximum Reduction in Width of the Prediction Interval for Each Aircraft

A/C	A/C Deleted	Reduction (%)
F-80	F2H-1	33.23
FH-1	F2H-1	31.06
F2H-1	F-102A	26.01
F7U-1	F2H-1	25.20
F-84E	F2H-1	25.17
F3D-1	F2H-1	33.22
F-86H	F2H-1	16.39
F4D-1	F2H-1	25.48
F3H-1N	F2H-1	25.59
F-102A	F2H-1	25.51
F-100D	F2H-1	31.31
FJ-4	F2H-1	29.53
F-104A	F9F-8	7.37
F11F-1	F2H-1	26.08
F-105B	F2H-1	28.79
F-101C	F2H-1	25.20
F-106B	F2H-1	25.49
F-5A	F2H-1	39.37
F-4J	F2H-1	32.78
F-111A	F2H-1	29.64
F-8E	F2H-1	31.78

BIBLIOGRAPHY

1. Allen, D. M. "Mean Square Error of Prediction as a Criterion for Selecting Variables." Technometrics 13: 469-476. 1971.

2. Baker, B. N. Improving Cost Estimating and Analysis in DoD and NASA, Ph.D. Thesis, George Washington University, Washington, D.C. 1972.

3. Cook, R. D. Detection of Influential Observations in Linear Regression. Technometrics 19: 15-18. 1977.

4. Cook, R. D. Influential Observations in Linear Regression. Journal of the American Statistical Association 74: 169-174. 1979.

5. Edwards, T. B. and Wallenius, K. T. On the Degree of Inflation of Measures of Fit Induced by Empirical Model Building. Unpublished Technical Report. Department of Mathematical Sciences. Clemson University. 1980.

6. Furnival, M. G., and R. W. Wilson, Jr. Regressions by Leaps and Bounds. Technometrics 16: 499-510. 1974.

7. Hoaglin, C. D., and R. E. Welsch. The Hat Matrix in Regression and ANOVA. The American Statistician 32: 17-22. 1978.

8. Korkotsides, A. S. and Wallenius, K. T. An Empirical Model Building Criterion Based on Prediction With Applications in Parametric Cost Estimation. Unpublished Technical Report. Department of Mathematical Sciences. Clemson University. 1980.

9. Large, J. P. Estimating Aircraft Turbine Engine Costs, Rand Corp. Report RM-6384/1-PR, Sept. 1970.

10. Levenson, G. S., et al. Cost Estimating Relationships for Aircraft Airframes, Rand Corp. Report R-761-PR, 1971.

11. Lindley, D. V. The Choice of Variables in Multiple Regression. Journal of Royal Statistical Society 30: 31-53. 1968.

12. Mahalanobis, P.C. On Tests and Measures of Group Divergence, J. Asiat. Soc. Beng. 26: 1930.

13. Mallows, C. L. Data Analysis in a Regression Context, Proceedings of Univ. of Kentucky Conference on Regression with a Large Number of Predictor Variables, Dept. of Statistics, University of Kentucky, Lexington, Ky.

14. Military Equipment Cost Analysis, RAND, prepared for the Office of Secretary of Defense (Systems Analysis), for Official use only, 1971.

15. Miller, Bruce M. and Sovereign, Michael G. Parametric Cost Estimating with Applications to Sonar Technology Naval Postgraduate School Technical Report NPS 55Z073091A. Monterey, California. 1973.

16. Rencher, A. C. and Pun, F. C. Inflation of R^2 in Best Subset Regression. Technometrics 22: 49-53. 1980.

17. Rondinelli, L. A. An Analysis of Cost Versus Performance Relationships for Phased Array Radars, RAND Corp., RM-5380-PR. 1967.

18. Sponsler, G. C., D. Gignoux, and N. N. Rubin. Parametric Cost Estimation of Fighter Aircraft. Technical Report. Columbia Research Corporation. Gaithersburg, Md. 1973.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER N121	2. GOVT ACCESSION NO. AD-A097 537	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Parametric Cost Methodology		5. TYPE OF REPORT & PERIOD COVERED Technical
		6. PERFORMING ORG. REPORT NUMBER TR #353
7. AUTHOR(s) A.S. Korkotsides, K. T. Wallenius		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-0451
9. PERFORMING ORGANIZATION NAME AND ADDRESS Clemson University, Dept. of Mathematical Sciences, Clemson, South Carolina 29631		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 047-202
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research, Code 434, Arlington, Va. 22217		12. REPORT DATE 17 Sept 1980
		13. NUMBER OF PAGES 22
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Parametric cost estimation, prediction, multiple regression, empirical model building, Mahalanobis distance, analogy.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report discusses the concept, advantages and disadvantages of parametric cost estimation and points to the need for explicitly incorporating degree analogy between the system whose cost is to be predicted and the historical data base. Mahalanobis distance is proposed as a measure of analogy and its properties studied. A new variable selection criteria is proposed which incorporates both fit and analogy. An example is given to illustrate its use.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: N121
2. GOVT ACCESSION NO.: AD-A097 537
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): Parametric Cost Methodology
5. TYPE OF REPORT & PERIOD COVERED: Technical
6. PERFORMING ORG. REPORT NUMBER: TR #353
7. AUTHOR(s): A.S. Korkotsides, K. T. Wallenius
8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0451
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Clemson University, Dept. of Mathematical Sciences, Clemson, South Carolina 29631
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR 047-202
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Code 434, Arlington, Va. 22217
12. REPORT DATE: 17 Sept 1980
13. NUMBER OF PAGES: 22
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:
16. DISTRIBUTION STATEMENT (of this Report): Approved for public release; distribution unlimited.
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report):
18. SUPPLEMENTARY NOTES:
19. KEY WORDS (Continue on reverse side if necessary and identify by block number): Parametric cost estimation, prediction, multiple regression, empirical model building, Mahalanobis distance, analogy.
20. ABSTRACT (Continue on reverse side if necessary and identify by block number):

This report discusses the concept, advantages and disadvantages of parametric cost estimation and points to the need for explicitly incorporating degree analogy between the system whose cost is to be predicted and the historical data base. Mahalanobis distance is proposed as a measure of analogy and its properties studied. A new variable selection criteria is proposed which incorporates both fit and analogy. An example is given to illustrate its use.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

LMED

8